Principali informazioni sull'insegnamento	
Titolo insegnamento	Data Mining
Corso di studio	Informatica e Tecnologie per la Produzione del Software
Crediti formativi	6
Denominazione inglese	Data Mining
Obbligo di frequenza	No
Lingua di erogazione	Italiano

Docente responsabile	Nome Cognome	Indirizzo Mail
	Donato Malerba	donato.malerba@uniba.it

Dettaglio crediti formativi	Ambito disciplinare	SSD	Crediti
	Informatico	ING-INF/05	6

Modalità di erogazione	
Periodo di erogazione	II semestre
Anno di corso	2018-2019
Modalità di erogazione	Lezioni frontali Esercitazioni in aula

Organizzazione della didattica	
Ore totali	150
Ore di corso	62
Ore di studio individuale	88

Calendario	
Inizio attività didattiche	01-03-2019
Fine attività didattiche	01-06-2019

Syllabus	
Prerequisiti	Conoscenze di basi di dati, algoritmica, statistica
Risultati di apprendimento previsti (declinare rispetto ai Descrittori di Dublino) (si raccomanda che siano coerenti con i risultati di apprendimento del CdS, compreso i risultati di apprendimento trasversali)	• *Conoscenza e capacità di comprensione* - Acquisizione di conoscenze relative agli algoritmi di data mining più noti in letteratura. - Comprensione delle scelte di algoritmi di data mining per specifici compiti. - Capacità di interpretazione dei risultati di un algoritmo di data mining. • *Conoscenza e capacità di comprensione applicate* - Capacità di realizzazione di un semplice progetto di scoperta di conoscenza in una collezione di dati mediante: - Utilizzo di strumenti per la selezione, pre-elaborazione e trasformazione dei dati, e per la validazione dei pattern estratti. - Utilizzo di strumenti di data mining per l'estrazione di

	conoscenza finalizzata a scopi predittivi e descrittivi in diversi contesti applicativi (aziendali e scientifici). • *Autonomia di giudizio* - Gli studenti sono in grado di apprezzare l'uso di algoritmi di data mining in processi di scoperta della conoscenza. - L'autonomia di giudizio viene acquista attraverso lo studio e l'interpretazione critica dei testi. - Il raggiungimento dell'adeguata autonomia è verificato attraverso le esercitazioni, che si tengono durante il corso, e con l'esame finale di profitto. • *Abilità comunicative* - Gli studenti sono in grado di esporre le tematiche incluse nel programma del corso mediante il lessico specifico della disciplina. • *Capacità di apprendere* - Gli studenti sono in grado di approfondire in autonomia le tematiche incluse nel programma del corso anche ricorrendo a risorse non direttamente coinvolte nella erogazione delle ore di lezione.
Contenuti di insegnamento	1. Scoperta di conoscenza nelle basi di dati: il processo La scoperta di conoscenza nelle basi di dati: definizione. Il processo della scoperta di conoscenza nelle basi di dati. Il processo CRISP-DM: business understanding, data understanding, data preparation, modelling, evaluation, deployment. 2. La classificazione Basata su alberi di decisione: Il modello ad albero. Test ai nodi interni e loro criteri di scelta. Il trattamento di attributi continui e mancanti. Complessità del problema di costruzione di alberi di decisione. La strategia "greedy" per la costruzione automatica di alberi di decisione. Limitazioni del modello ad albero. Trasformare alberi di decisione in insiemi di regole. Il rasoio di Occam. Costruire alberi di decisione della giusta dimensione: la semplificazione (pruning). Basata su insiemi di regole: Costruzione automatica di descrizioni di un concetto a partire da esempi. Strutturare lo spazio delle ipotesi con ordini di generalità. L'algoritmo Find-S per la ricerca di ipotesi massimamente specifiche. Limiti di Find-S. Lo spazio delle versioni e l'algoritmo di eliminazione dei candidati. Classificare con spazi delle versioni. Trattare dati rumorosi. Il ruolo della polarizzazione (bias) induttiva nella costruzione automatica di ipotesi. Apprendere definizioni disgiuntive di concetti: la strategia di di copertura sequenziale (sequential covering o separate-and-conquer). Il passo della 'conquista' nella copertura sequenziale: la ricerca a

	raggio (beam search). Complessità computazionale di algoritmi fondamentali per l'apprendimento di classificatori a regole. Costruzione automatica di descrizioni di più concetti a partire da esempi. Basato su teorema di Bayes: Il teorema di Bayes e nozioni di base correlate. Apprendimento Maximum-A-Posteriori forza bruta. Ipotesi MAP e sistemi di apprendimento consistenti. Il principio MDL (minimum description length). Classificatori ottimali di Bayes. Algoritmo di Gibbs. Il classificatore di naive Bayes e la stima delle probabilità. La classificazione di testi basata su classificatori naive Bayes. 3. La regressione parametrica e non parametrica Modelli di regressione con errore additivo. Modelli di regressione (lineare) semplice: stima dei parametri con il metodo dei minimi quadrati, aspetti inferenziali del modello di regressione semplice, diagnostiche grafiche del comportamento dei residui. Modelli di regressione lineare semplice con regressore qualitativo: stima dei parametri con il metodo dei minimi quadrati, aspetti inferenziali del modello di regressione semplice, diagnostiche grafiche del comportamento dei residui. Estensioni: regressione polinomiale e regressione lineare multipla. La notazione matriciale. Funzioni costanti a tratti per la regressione semplice. I modelli non parametrici: alberi di regressione. Combinare modelli parametrici e non: gli alberi dei modelli. L'algoritmo SMOTI per l'apprendimento di alberi di modelli. 4. Le associazioni di variabili Il caso di due variabili - Correlazioni di variabili quantitative: coefficiente di correlazione (parziale) di Pearson. Coefficienti di correlazione e modelli di regressione. Associazioni di variabili qualitative: lambda, coefficiente di Spearman.. Il caso di più variabili – I limiti dei modelli log-lineari. Le regole di associazione: le configurazioni (pattern) frequenti, supporto di una configurazione, confidenza di una regola di associazione, le regole di associazione forti, il metodo Apriori per la generazione di regole di associazione forti. 5. Tecniche avanzate di data mining Data Mining Multi-Relazionale: assunzione di tabella singola, dati distribuiti su più tabelle, pattern relazionali, come promuovere gli algoritmi proposizionali all'analisi di dati relazionali. Uno studio di caso: SPADA. La classificazione relazionale con FOIL, Progol e TILDE.

Programma	
Testi di riferimento	T. Mitchell Machine Learning

	Morgan Kaufmann, 1997 Capitoli: 2-3-6 Richard J. Roiger, Michael W. Geatz. Introduzione al Data Mining. McGraw-Hill, 2003 Capitoli: 1-2-3-4-5-6-8-9 A. Azzalini, B. Scarpa Analisi dei dati e data mining Sprinter, 2004 Capitoli: 1-4-5-6 Articoli scientifici selezionati e copia delle trasparenze proiettate durante le lezioni e durante le esercitazioni in laboratorio sono disponibili sul sito: http://www.di.uniba.it/~malerba/courses/bcdm.htm
Note ai testi di riferimento	Testo adottato
Metodi didattici	- Lezioni frontali condotte con l'ausilio di supporti didattici (slide) - Esercitazioni
Metodi di valutazione (indicare almeno la tipologia scritto, orale, altro)	- Prova scritta sulla parte teorica - Svolgimento di un progetto di scoperta di conoscenza nei dati mediante l'applicazione di algoritmi di data mining. - La prova scritta è propedeutica alla presentazione del progetto.
Criteri di valutazione (per ogni risultato di apprendimento atteso su indicato, descrivere cosa ci si aspetta lo studente conosca o sia in grado di fare e a quale livello al fine di dimostrare che un risultato di apprendimento è stato raggiunto e a quale livello)	Nella prova scritta, lo studente dev'essere in grado di esporre, in modo critico, i concetti appresi relativi al processo di scoperta della conoscenza. Dev'essere altresì capace di affrontare semplici esercizi di data mining. Nel progetto lo studente deve dimostrare capacità di analisi dei dati, di applicazione di algoritmi di data mining e di comprensione dei risultati ottenuti, in un ciclo finalizzato al miglioramento delle prestazioni. Gli strumenti che si possono all'uopo utilizzare sono quelli illustrati durante le ore di esercitazione/laboratorio.
Altro	